scientific reports

Check for updates

OPEN

A mode of action protein based approach that characterizes the relationships among most major diseases

Hongyi Zhou, Brice Edelman & Jeffrey Skolnick✉

Disease classification is important for understanding disease commonalities on both the phenotypical and molecular levels. Based on predicted disease mode of action (MOA) proteins, our algorithm PICMOA (Pan-disease Classification in Mode of Action Protein Space) classifies 3526 diseases across 20 clinically classified classifications (ICD10-CM major classifications). At the top level, all diseases can be classified into "infectious" and "non-infectious" diseases. Non-infectious diseases are classified into 9 classes. To demonstrate the validity of the classifications, for common pathways predicted based on MOA proteins, 77% of the top 10 most frequent pathways have literature evidence of association to their respective disease classes/subclasses. These results indicate that PICMOA will be useful for understanding common disease mechanisms and facilitating the development of drugs for a class of diseases, rather than a single disease. The MOA proteins, molecular functions, pathways for classes, and individual diseases are available at https://sites.gatech.edu/cssb/PICMOA/.

Different diseases do not have fully independent underlying molecular causes; rather, they are densely connected as demonstrated by comorbidity studies, where apparently disparate diseases cooccur more often than would be expected at random[1–3]. These dense connections can be attributed to their shared molecular profiles. However, in practice it has proven to be difficult to create a dendrogram that characterizes the interrelationships of all diseases[2]. Furthermore, even patients with the same clinical classification (e.g. the ICD10-CM major classifications[4]) might respond to different treatments[5,6]. These differences are often due to disparate underlying molecular causes despite the same clinical diagnosis[5,7]. A disease's molecular profile could be described by differential gene expression data, DNA methylation patterns, copy number variation, missense and nonsense mutations, or a single, few or perhaps many protein targets[6]. Here, we focus on the malfunctioning of the driver proteins that possibly cause a set of diseases; these unique sets of mode of action (MOA) proteins of each patient are likely responsible for personalized treatment outcome. By examining the intersection of the MOA protein profiles of comorbid diseases, one can narrow down their common cause.

Classifying diseases into appropriate molecular classes could suggest repurposed uses of existing treatments or drive new therapeutic strategies. For example using k-means classification, a recent work that classified ~ 10,000 cancer patients of 32 types using differential gene expression data based on mRNA expression[7] found that multiple cancer types occur within a given molecular cancer classification. Many years of efforts by the TCGA project made it possible to systematically analyze cancer diseases/patients[6,8]. Given this level of effort, it is quite unlikely that a similar project will be undertaken for all other human complex diseases in the near future. Even if this were possible, for pan disease classification, the basis vector (e.g. the top 2000 genes of the most differentially expressed mRNA data across 32 types of cancers) of the cancer classification provided in Ref.[7] is not readily available for most diseases. Moreover, normalization of gene expression levels across complex diseases is also challenging.

As pointed out by the earlier work of[9], phenotype clusters can be attributed to gene expression patterns, coevolution, or gene ontology. In practice, except for[7] that classified 32 cancer types using differential gene expression data (called pan-cancer classification even though there are more than 500 types of cancers[10]), many other approaches classify a relatively small number of diseases/cancers: Tibshirani et al. classified small round blue cell tumors and leukemias from gene expression profiling[11]; Reinius et al. utilized DNA methylation for classifying cell lineage in complex inflammatory diseases[12]; Copy number variation was able to distinguish healthy from diseased tissues[13] and was applied for subclasses of Pancreatic Cancer[14]. Specific gene mutations

Center for the Study of Systems Biology, School of Biological Sciences, Georgia Institute of Technology, 950 Atlantic Drive, N.W., Atlanta, GA 30332, USA. ✉email: skolnick@gatech.edu

can also relate diseases such as those in the Online Mendelian Inheritance in Man (OMIM) database[15]. Due to data availability, these studies are limited to specific types of diseases, e.g. cancer or inflammatory diseases. The OMIM database is limited to connections among Mendelian diseases that are possibly linked to cancers[16]. In short, these methods are difficult to apply to the large numbers of diseases across the 20 types of major clinical classifications[17] required for the examination of their more global inter-relationships. Another shortcoming of these studies is that the input genomic data does not distinguish efficacious drug targets (drivers) from downstream unimportant targets; both are convoluted in the expression data.

To generally address these issues, we explore a novel approach that uses MOA proteins as the key molecular descriptors. These MOA proteins are then associated with a set of diseases. However, they are not necessarily the disease drivers. A given MOA protein can contribute to the disease but if it is downstream/upstream of the key disease driver proteins, they could be responsible for differential response—e.g. in the brain, one protein exacerbates Alzheimer's disease, while in the thyroid, another protein exacerbates hyperthyroidism but does not cause it. In contrast, knowing the driver proteins could suggest a common treatment of all such diseases, much like a broad spectrum antibiotic can treat infections caused by different types of bacteria.

How can these key driver proteins be determined? In our previously developed LeMeDISCO[2] algorithm, we derived the set of common protein drivers among 3526 diseases. These proteins are the population averaged profile of the disease. These MOA driver proteins were used to predict disease comorbidity and performed essentially as well as a symptom based approach that lacks any molecular explanation[1]. For a benchmark set of 2630 comorbid disease pairs, 71% are correctly classified by LeMeDISCO as being comorbid. Here, for an individual disease, its molecular profile is built from predicted MOA proteins provided by LeMeDISCO[2]. We emphasize that only proteins which are expressed in the given tissue where the disease occurs are considered[2]. The aim of this work is to classify a collection of 3526 complex diseases into molecular subtypes and provide their respective MOA proteins, molecular functions and pathways.

Results

Disease enriched mode of action proteins

The top 50 proteins and disease enriched MOA proteins assigned by their normalized p-value weighted frequencies are given in Fig. 1, with the top 2000 MOA proteins listed in Supplementary Table S1. These proteins are prioritized by their likelihood of being a disease driver. The inset shows that the top 50 MOA proteins are involved in over 70% of diseases. Many are associated with over 1000 diseases (around 1/3 of the 3526 diseases). The high frequency of common proteins that connect one disease to many others results in a very

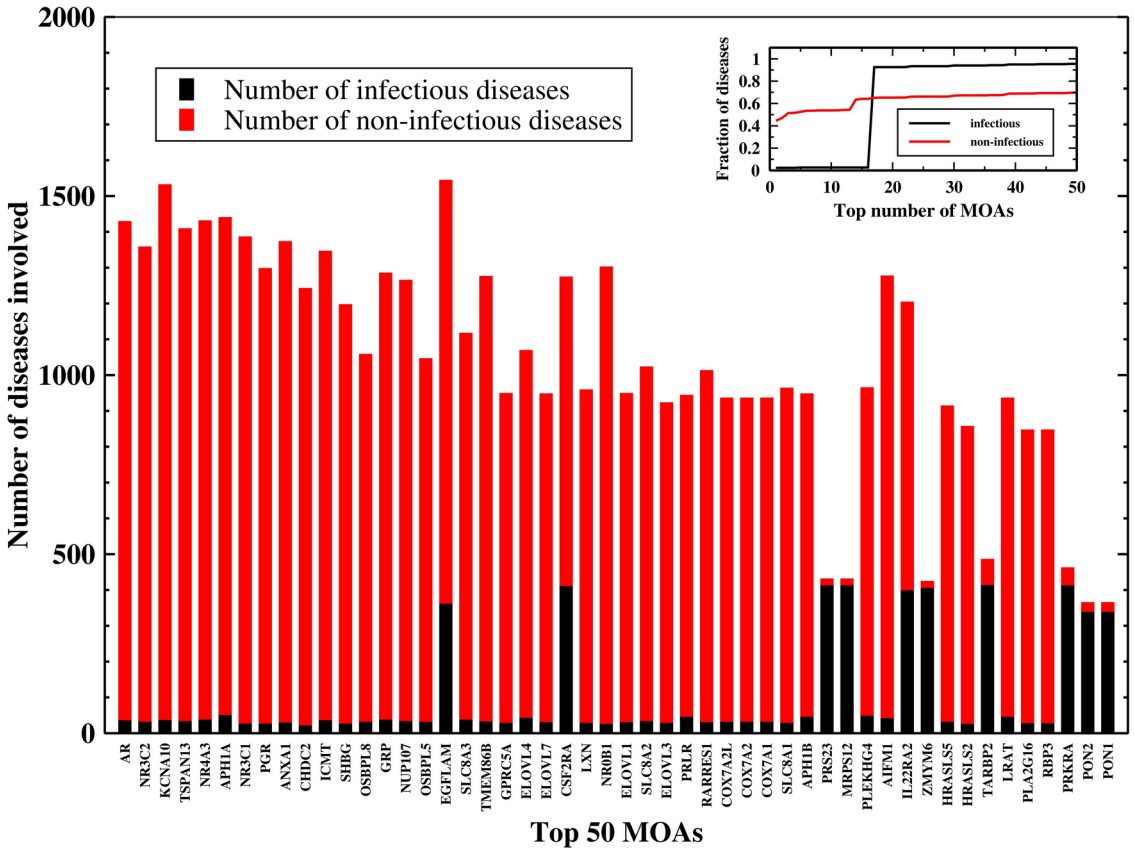

Fig. 1. The top 50 proteins and their disease frequencies for 3526 diseases. Inset: cumulative fraction of diseases vs. top number of MOA proteins.

dense disease network. Next, we provide some details about the top 5 genes: AR, NR3C2, KCNA10, TSPAN13, NR4A3. They are associated with 1428, 1357, 1531, 1408, and 1430 of the 3526 diseases, respectively. Three (AR, NR3C2, NR4A3) belong to the nuclear receptor family and regulate other genes. All three have DNA binding sites, especially two zinc finger domains[18]. The regulatory functions and ubiquity of well-studied zinc fingers in these proteins may explain their frequent presentation as disease MOA proteins[19]. KCNA10, which is expressed in many tissues, is a voltage-gated potassium channel subunit and is known to be associated with Developmental and Epileptic Encephalopathy 32 and Episodic Ataxia, Type 1 and may be associated with other neurological disorders[20]. TSPAN13 is a member of the transmembrane 4 tetraspanin superfamily. These proteins mediate signal transduction events that play a role in the regulation of cell development, activation, growth, and motility[21]. We found that our top 2000 MOA proteins have 288 overlapped ones with the top 2000 most differentially expressed proteins of cancer patients found in the pan-cancer classification work of Ref.[7]. The overlap is significant with a p-value of 9.05×10^{-6}. This again indicates that our predicted MOA proteins are indeed biologically relevant to at least a subset of diseased patients.

Pan-disease classification

Using the above top 2000 disease enriched MOA proteins (see Table S1) as the basis vector for each of 3526 diseases defined by the Human Disease Ontology[10] and analyzed in MEDICASCY for indication predictions[22], we classify disease interrelationships by a k-means approach (see "Methods"). The Silhouette Coefficient[23] (SC) indicates that k = 2 (SC = 0.657, the next best SC = 0.510 at k = 11) is the optimal number of clusters. The statistics of the resulting pan-disease classification is given in Table 1. Cluster 1 (C1) contains 416 diseases, with the majority, 305 (75%), being Certain infectious and parasitic diseases. Cluster 2 (C2) contains 3110 diseases; only 31 are members of the Certain infectious and parasitic diseases class. This indicates that what was traditionally clinically classified as infectious diseases and non-infectious disease are the main, but nonexclusive constituents of the two classes, respectively. Rather, diseases within a given class share common disease driving mechanisms.

Stability and sensitivity of clustering

To show that the above procedure is stable, we performed a random test by changing the initial partition of the clusters by using 100 different random seeds. The Silhouette Coefficient SC = 0.657 at k = 2 is identical for all 100 random seeds. In addition, we also recorded the ratio of sum of distance squares (SS) between clusters vs. total sum of distance squares (between_SS/total_SS where total_SS = between_SS + within_SS, within_SS = sum of SS within clusters). A naïve cluster number equal to the number of data points will give this ratio a value of 1, within_SS = 0 for this situation. This value (= 55% at k = 2) is also identical for all 100 iterations. Knowing that k = 1 will have between_SS/total_SS = 0, a 55% at k = 2 is an "elbow" point according to the Elbow method for determining the optimal k value[24]. An elbow point is found where the increase between_SS/total_SS ratio goes down when an additional number of clusters is added. Here, increasing k from 2 to 3 increases the between_SS/total_SS ratio by 14%: from 55% to 69%. This 14% increase is much smaller than the increase 55% from k = 1 to k = 2. Thus k = 2 is also optimal cluster number by the Elbow method. Its corresponding metrics SC and between_SS/total_SS are stable to random perturbation. Next, for k = 2 clustering, we perform 100 tests of random shuffling the disease-cluster assignments and calculate SC and between_SS/total_SS. The average SC and between_SS/total_SS are − 0.001 and 3.0% with standard deviations of 0.018 and 2.2%. These results indicate that both SC and between_SS/total_SS are sensitive to correct disease-cluster assignments. Complete random assignments result in worse values.

Differences in the MOA proteins between infectious and non-infectious diseases

In Table 2, we list the top 20 most frequent MOA proteins among diseases in the infectious dominated class, C1, and non-infectious dominated disease class, C2. None overlap. For C1, almost all 416 diseases involve the same top 20 MOA proteins. 17 of these top 20 MOA proteins in class C1 have supporting evidence for association with infectious diseases (see Table 2). In contrast in class C2, less than half (1494/3110 = ~48%) have the same top first MOA proteins. This indicates that C1 is homogeneous while C2 is more divergent.

High level classification of infectious diseases

Among the 111 diseases in class C1 not clinically classified as infectious diseases are 18 *Diseases of the eye and adnexa*, 15 *Diseases of the digestive system*, 12 *Neoplasms*, 10 *Diseases of the musculoskeletal system and connective tissue*, and 9 *Diseases of the respiratory system*. One example is constipation clinically assigned in *Diseases of the digestive system*. Does this make sense? Intestinal bacteria and chronic constipation are correlated[25]. Similarly, colorectal cancer of *Neoplasms* is infection associated[26]. Diabetic angiopathy of *metabolic diseases* is associated with COVID-19 infection[27]. We also classify meningitis and bacterial meningitis (*Diseases of the nervous system*), and respiratory and kidney failure into C1. The classification of these diseases into C1 indicates that they have similar molecular mechanisms as infectious diseases and might be prevented/treated with a similar strategy, for example, by vaccination. All diseases classified at this level are found in Table S2.

High level classification of non-infectious diseases

The 3110 diseases of class C2 are > 99% non-infectious diseases and have divergent MOA proteins. Moreover, class C2 contains the majority of and the most diverse clinical classifications of the total number of diseases considered (3110 of 3526). In contrast, class C1 mainly involves infectious diseases. As seen in our pathway analysis, the top 20 most frequent pathways are present in > 95% of the diseases in class C1, whereas they are present in only around 1/3 of the diseases in class C2. Thus, C1 is more homogeneous and C2 is more divergent. Therefore, we further classify C2 with the same procedure as global pan-disease classification. The optimal SC score is 0.487 at k = 9. For level 2 classification, we denote the resulting 9 subclasses as C2-j, where j = 1,2,…9

Number of diseases	Clinical classification
cluster C1: 416 diseases	
305	Certain_infectious_and_parasitic_diseases
18	Diseases_of_the_eye_and_adnexa
15	Diseases_of_the_digestive_system
12	Neoplasms
10	Diseases_of_the_musculoskeletal_system_and_connective_tissue
9	Diseases_of_the_respiratory_system
8	Diseases_of_the_skin_and_subcutaneous_tissue
7	Diseases_of_the_genitourinary_system
7	Diseases_of_the_circulatory_system
6	Diseases_of_the_nervous_system
5	Congenital_malformations,_deformations_and_chromosomal_abnormalities
4	Endocrine,_nutritional_and_metabolic_diseases
3	Factors_influencing_health_status_and_contact_with_health_services
3	Diseases_of_the_ear_and_mastoid_process
2	Pregnancy,_childbirth_and_the_puerperium
1	Diseases_of_the_blood_and_blood-forming_organs_and_certain_disorders_involving_the_immune_mechanism
1	Certain_conditions_originating_in_the_perinatal_period
cluster C2: 3110 diseases	
547	Neoplasms
527	Diseases_of_the_eye_and_adnexa
430	Diseases_of_the_nervous_system
402	Endocrine,_nutritional_and_metabolic_diseases
224	Diseases_of_the_circulatory_system
196	Diseases_of_the_digestive_system
118	Mental_and_behavioural_disorders
106	Diseases_of_the_respiratory_system
104	Diseases_of_the_genitourinary_system
99	Diseases_of_the_blood_and_blood-forming_organs_and_certain_disorders_involving_the_immune_mechanism
91	Diseases_of_the_skin_and_subcutaneous_tissue
86	Diseases_of_the_musculoskeletal_system_and_connective_tissue
80	Congenital_malformations,_deformations_and_chromosomal_abnormalities
31	Certain_infectious_and_parasitic_diseases
17	Symptoms,_signs_and_abnormal_clinical_and_laboratory_findings,_not_elsewhere_classified
17	Diseases_of_the_ear_and_mastoid_process
12	Pregnancy,_childbirth_and_the_puerperium
9	Certain_conditions_originating_in_the_perinatal_period
8	Injury,_poisoning_and_certain_other_consequences_of_external_causes
6	Factors_influencing_health_status_and_contact_with_health_services

Table 1. Statistics of pan-disease classification in disease classes C1 and C2.

which are the output labels of the k-means method. A summary is given in Fig. 2. The full list of subclasses, enriched GO molecular functions, and Reactome pathways are presented in Tables S3 (subclasses), S4 (GO functions) and S5 (pathways), respectively. Five of nine subclasses are dominated by a single clinical classification (>70%): C2-1 by digestive, C2-2 by metabolic, C2-4 purely by neoplasms, C2-6 by eye, and C2-8 by neoplasm diseases. Although C2-3, C2-5, C2-7 are diverse in their clinical disease assignments, their topmost frequent GO molecular functions and Reactome pathways are present in almost all member diseases. Subclass C2-9 contains 1405 diseases and based on its top frequent GO molecular functions (Table S4) and Reactome pathways (Table S5), this class still contains a diverse set of disease types.

We then classify the 1405 diseases in C2-9 into k = 17 level 3 subclasses denoted as C2-9-1, C2-9-2,…, C2-9-17 (again, the last digit is the output label of k-means clustering). A summary of this third level classification is given in Fig. 3. The full list of third level subclasses and the enriched GO molecular functions at this level are given in Tables S6 (subclasses), S7 (GO functions), and S8 (pathways), respectively. There are four subclasses (C2-9-2:5/7, C2-9-12:117/124, C2-9-15:8/11, C2-9-16:13/16) dominated by *Neoplasms*; two subclasses (C2-9-1:6/6, C2-9-13:3/4) dominated by *Diseases of the circulatory system*; one subclass (C2-9-6:58/98) associated with *Diseases of the respiratory system*; one subclass (C2-9-7:261/265) associated with *Diseases of the nervous system*; and one subclass (C2-9-9:18/20) associated with *Congenital malformations, deformations and chromosomal abnormalities*.

Gene	Disease frequency	Literature evidence
416 diseases of class C1		
TEX101	416	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7404878/
REEP1	416	https://www.frontiersin.org/articles/10.3389/fvets.2021.779323/full
PROK2	416	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3016599/
GNRH2	416	https://journals.asm.org/doi/pdf/10.1128/msystems.00555-23
FOXF1	416	https://journals.lww.com/md-journal/fulltext/2021/04090/de_novo_mutation_of_foxf1_causes_alveolar.67.aspx
DISC1	416	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3335463/
ACRV1	416	https://www.genecards.org/cgi-bin/carddisp.pl?gene=ACRV1
TSSC4	415	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2904600/
SIGLEC15	415	https://jbiomedsci.biomedcentral.com/articles/10.1186/s12929-019-0610-1
PRCC	415	
PLAUR	415	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9785175/
NTNG1	415	https://www.researchgate.net/publication/341958756_SARS-CoV-2_orf1b_gene_sequence_in_the_NTNG1_gene_on_human_chromosome_1
MDFIC	415	https://www.uniprot.org/citations/16260749
LRTOMT	415	
LRRC27	415	https://academic.oup.com/jid/article/217/7/1044/4782485
LRRC17	415	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4077721/
ESM1	415	
CRTC3	415	
CRIPAK	415	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4820667/
CDHR4	415	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7854084/
3110 diseases of class C2		
KCNA10	1494	
NR4A3	1392	
AR	1392	
APH1A	1388	
TSPAN13	1374	
NR3C1	1358	
ANXA1	1342	
NR3C2	1325	
ICMT	1309	
NR0B1	1275	
PGR	1270	
GRP	1246	
TMEM86B	1242	
AIFM1	1234	
NUP107	1230	
CHDC2	1219	
EGFLAM	1181	
SHBG	1169	
SLC8A3	1078	
FAM26E	1042	

Table 2. Top 20 most frequent MOA proteins in disease classes C1 and C2.

Relationships between diseases and molecular classes

To have a complete picture of all disease-disease relationships, we next plot the dendrogram of all 3526 diseases using their Euclidean distance matrix between their MOA protein vectors as assessed by the minimum evolution MEGA method[28]. Due to space limitations, the full tree is shown in Supplementary Fig. S1. For each disease, the diagram shows its close neighbors. For example, searching "constipation" will find as its neighbors "irritable bowel syndrome", "pericardial effusion", etc. An unusual neighbor is "brain disease" of *Diseases of the nervous system*. Both "constipation" and "brain disease" are classified in C1 dominated by infectious diseases. A Sankey diagram showing the flow of clinical classifications to molecular classes is given in Fig. 4. The majority of infectious diseases go to class C1; eye disease to C2-6; metabolic diseases to C2-2; circulatory diseases to C2-9-4; digestive diseases to C2-1; and nervous diseases to C2-9-7. Neoplasms are split into many classes without a dominated one. These relationships help us understand the disease cause/comorbidity of a given disease.

For class relationships, the centroid vectors of the 26 molecular disease classes were used to calculate the Euclidean distance matrix between them with the resulting dendrogram in Fig. 5 describing their distance/

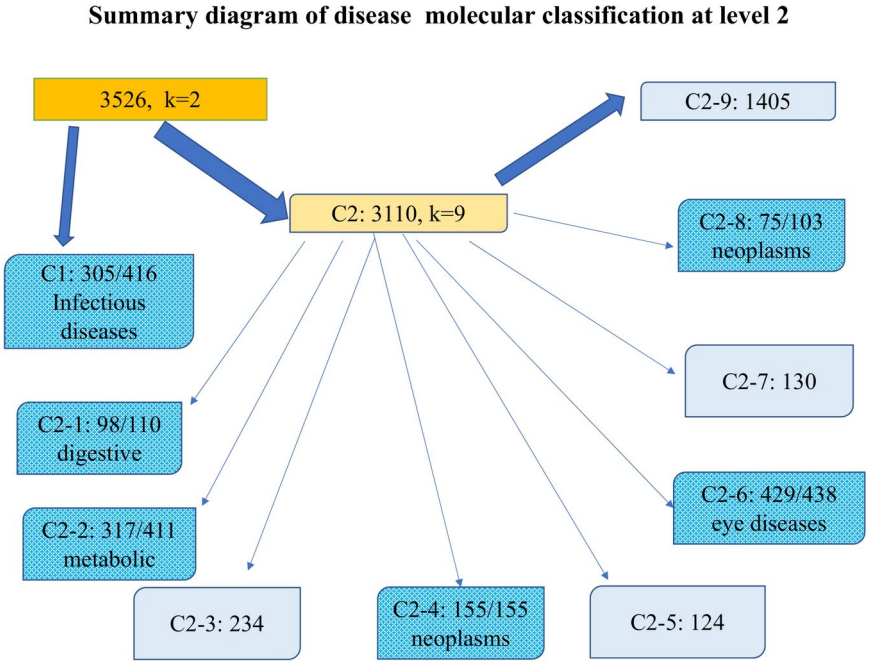

Fig. 2. Summary of disease molecular classification at level 2. Dark (light) blue colored subclasses are dominated (70%) by a single (multiple) clinical classification(s). The texts in the boxes are the most frequent clinical classifications. The number in each box is the total number of diseases or number of diseases having the most frequent clinical classification/total number diseases in that class.

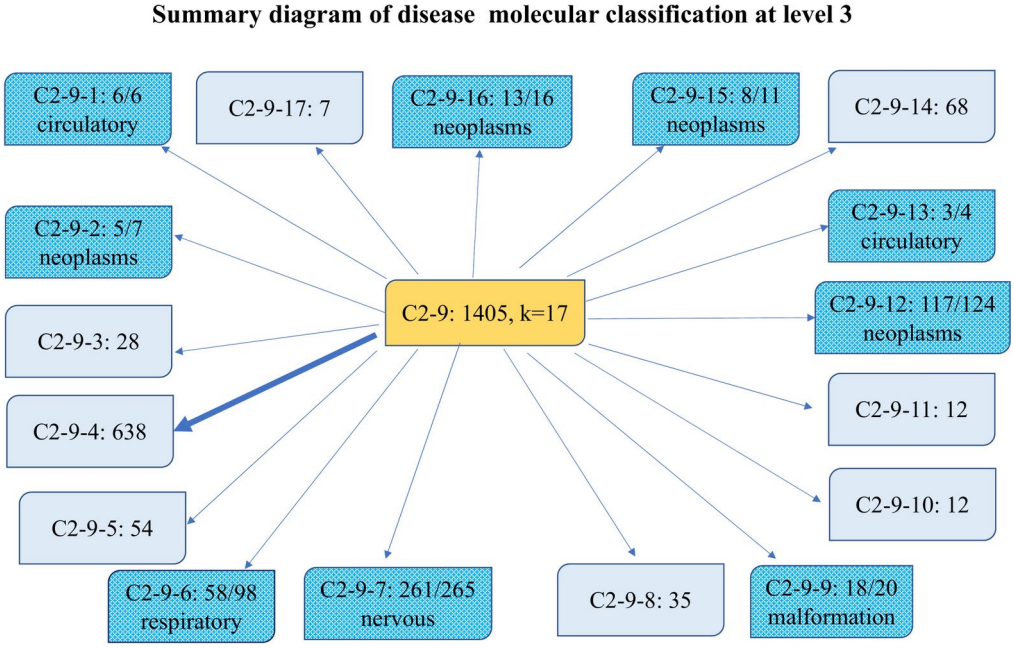

Fig. 3. Summary of disease molecular classification at level 3. Dark (light) blue subclasses are dominated (>70%) by a single (multiple) clinical classification(s). The texts in the boxes are the most frequent clinical classification. The number in each box is the total number of diseases or number of diseases having the most frequent clinical classification/total number diseases in that class.

closeness in molecular space. Infectious disease class C1 is at the root of the tree and most distant from the rest. A small class C2-9-13 with 3/4 circulatory diseases (*varicose veins, anterolateral myocardial infarction, anteroseptal myocardial infarction*) is distant from the others. This also means that C2-9-13 is the closest class to infectious disease class C1. In fact, we predicted that of 95 significant pathways for C2-9-13, 92 overlapped

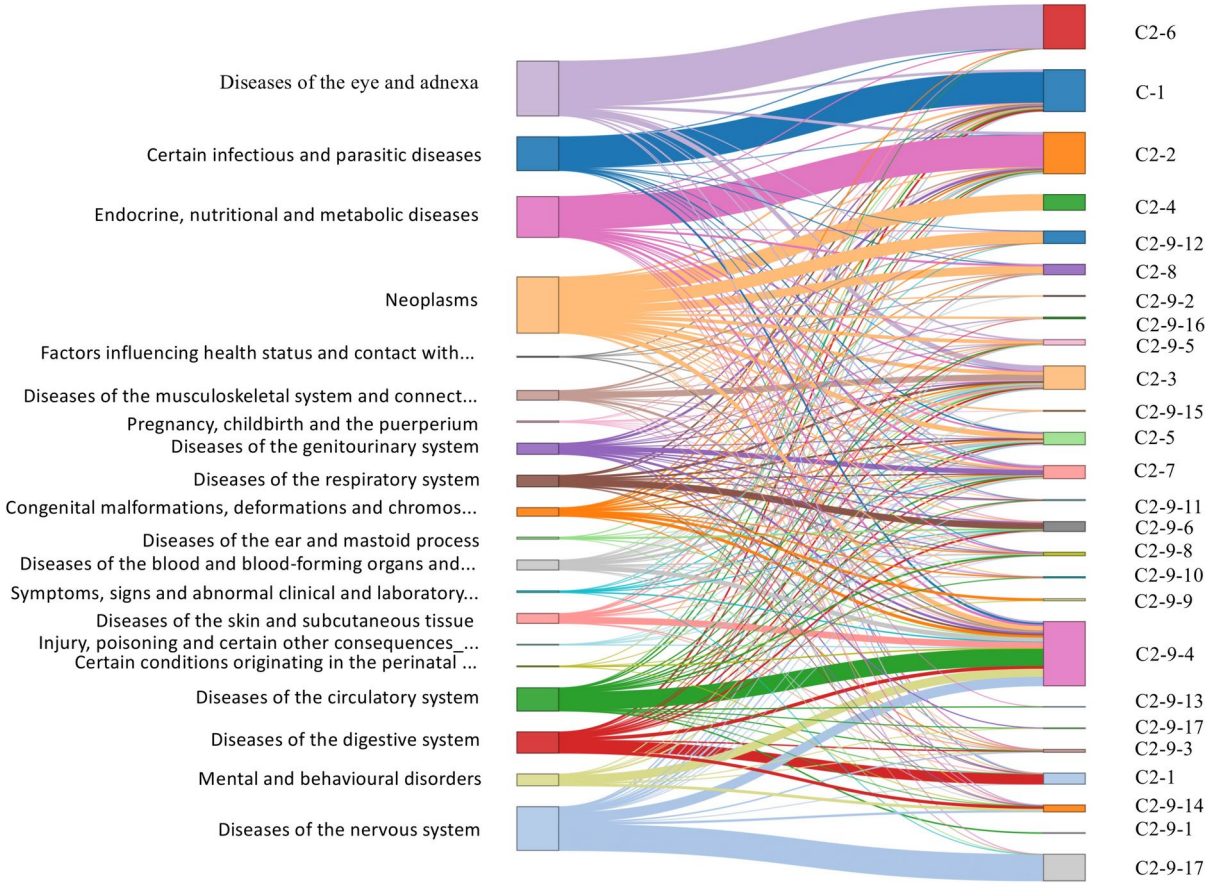

Fig. 4. Sankey diagram that maps the clinical classifications of diseases into their molecular classes (The underlying data is found at https://github.com/hzhou3ga/picmoa/).

with those of C1. For example, *Immune System, Innate Immune System, Cytokine-cytokine receptor interaction, Cytokine Signaling in Immune system, and Jak-STAT signaling pathway* are among the top 20 most frequent pathways of both classes. Figure 5 also shows that class C2-7, a diverse (i.e., not dominated by a single clinical classification) class, includes many common diseases such as *Alzheimer's disease, asthma, central nervous system disease, Crohn's disease, coronary artery disease, opiate dependence, prostate cancer*, is close to C2-6 dominated by eye disease. We will analyze their molecular connections in detail below. Note that this clustering simultaneously provides the sets of comorbid diseases.

Molecular functions and pathways of classes

Next, we examine the molecular functions and pathways enriched in the different classes. In each class, we use the MOA protein component vector of each disease to search for their GO molecular functions[29,30] and Reactome pathways[31] and then count the frequencies of each function and pathways. The top 20 most frequent GO functions and Reactome pathways for C1 and C2 are compiled in Tables 3 and 4, respectively. The complete frequency list is given in Supplementary Tables S9 (GO functions) and S10 (pathways). Full lists of molecular classes, GO functions, and pathways for individual diseases as well as molecular classes are available at https://sites.gatech.edu/cssb/PICMOA/.

From Table 3, the top 20 most frequent GO functions have no overlap between the two classes. In Class C1, almost all 416 diseases have the same 20 enriched GO functions. Literature mining found 15 articles that support the prediction that the top 20 predicted GO functions of C1 are associated with infection (see Table S9); whether the other predictions are correct or not is uncertain. For the non-infectious class C2, the top GO function, "steroid binding" has a frequency of 1096, i.e. only around 1/3 of the 3110 diseases have this GO function enriched. Thus, class C2, though well separated from C1, has very more diverse molecular functions.

Table 4 shows a similar trend as in Table 3 that the top 20 pathways are present in > 95% of the diseases in class C1, whereas they are present in around 1/3 of diseases in class C2. Thus, the top pathways in C1 are more specific to infectious diseases. There are no overlapping pathways between the two sets of top 20 pathways. Within the top 20 most frequent pathways of C1, many are related to immune system/signaling. A literature search shows that 19 (all but *Defective B4GALT1 causes B4GALT1-CDG (CDG-2d)*) of the 20 pathways have evidence of association with infectious diseases (see Table S10). For example, the top pathway, *Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell*, is associated with the difference between Sepsis and normal cell samples[32]. The second ranked pathway, *immune system*, has long been known to be linked with infectious diseases[33].

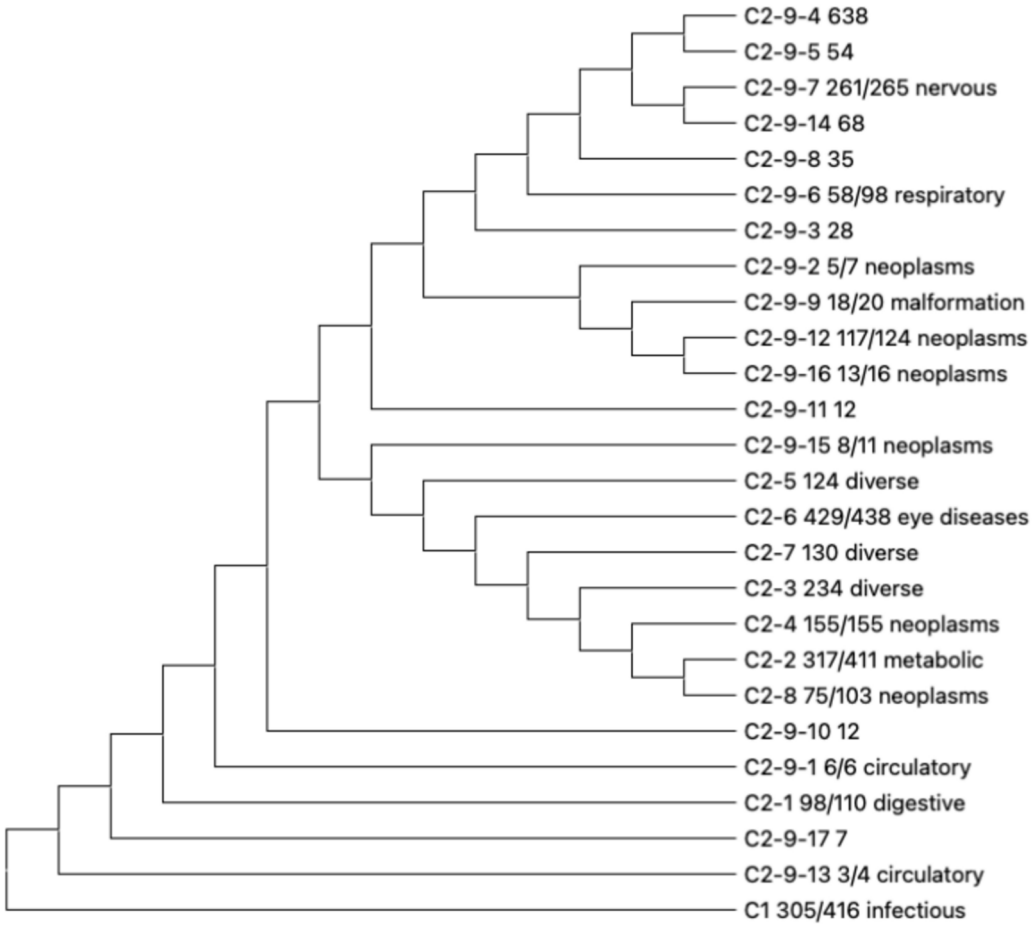

Fig. 5. Dendrogram showing the closeness of the 26 disease classes.

The third, the *Adaptive Immune System*, is involved with the well-established antigen-specific responses of T and B lymphocytes to infection[34,35]. The fourth, *Signaling by Interleukins* and the *adaptive immune system* and *Phagosome*, is enriched among COVID-19 patients[35]. The fifth, *Innate immune system,* is the first pathway that initiates antiviral response[36].

Pathway analysis of non-infectious diseases

Next, we examine some examples of the non-infectious disease subclasses for their common enriched pathways (top pathways present in more than 95% of their member diseases). Experimental/bioinformatics evidence for one member of the subclass is likely transferable to all members. We searched for literature evidence for the top 10 most frequent pathways; the results are compiled in Supplementary Tables S5 and S8. Literature evidence is considered correct if we find one of the class members is associated with the pathway. Summaries of the number of diseases, top clinical classification, frequencies of the top pathway, and GO function of each class are given in Supplementary Table S11. We detail some results below.

Class C2-1 is dominated by digestive diseases (98/110) and contains half of all digestive diseases. Its top 20 most frequent pathways have 8 overlaps with the top 20 with infectious disease class C1. This indicates that this subclass of diseases is close to the infectious diseases in terms of molecular mechanisms that mainly involve immune system/signaling pathways. In fact, 15 *Diseases of the digestive system* are classified in the infectious disease class C1. The top pathway is identical to that of class C1: *Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell*. This pathway is also associated with *tooth agenesis*[37]. The *Immune System* pathway is associated with *colitis*[38]. For 7 of the top 10 pathways of C2-1, we found supporting literature (see Table S5).

Class C2-2 has 411 diseases with 317 *Endocrine, nutritional and metabolic diseases* including *type 1 diabetes mellitus (T1DM)*. 21 *Diseases of the eye and adnexa* are also included in this class. Among its top 10 frequent pathways are four fatty acid related pathways: *Synthesis of very long-chain fatty acyl-CoAs*, *Fatty acid biosynthesis elongation endoplasmic reticulum*, *Fatty acid elongation* and *Fatty Acyl-CoA Biosynthesis*. Recent evidence shows that free fatty acids are associated with the development of islet autoimmunity in T1DM[39].

Class C2-4 contains 155 pure *Neoplasms diseases* (including *colon cancer*) out of 547 cancers from class C2. The top 5 pathways of this subset of cancers are *Olfactory Signaling Pathway*[40], *Signaling by GPCR*[41], *Signal*

Frequency	GO function
416 diseases of class C1	
416	GO:0038023 signaling receptor activity
416	GO:0004888 transmembrane signaling receptor activity
415	GO:0030246 carbohydrate binding
414	GO:0033691 sialic acid binding
412	GO:0030021 extracellular matrix structural constituent conferring compression resistance
410	GO:0016936 galactoside binding
410	GO:0004896 cytokine receptor activity
408	GO:0008201 heparin binding
408	GO:0005225 volume-sensitive anion channel activity
407	GO:0016019 peptidoglycan immune receptor activity
407	GO:0008745 *N*-acetylmuramoyl-L-alanine amidase activity
405	GO:0008083 growth factor activity
405	GO:0005537 mannose binding
405	GO:0005104 fibroblast growth factor receptor binding
404	GO:0042834 peptidoglycan binding
403	GO:0070492 oligosaccharide binding
402	GO:0061809 NAD + nucleotidase, cyclic ADP-ribose generating
402	GO:0050135 NAD(P) + nucleosidase activity
402	GO:0005111 type 2 fibroblast growth factor receptor binding
402	GO:0005105 type 1 fibroblast growth factor receptor binding
3110 diseases of class C2	
1096	GO:0005496 steroid binding
1031	GO:0004930 G protein-coupled receptor activity
993	GO:0003707 nuclear steroid receptor activity
953	GO:0004984 olfactory receptor activity
908	GO:0102756 very-long-chain 3-ketoacyl-CoA synthase activity
908	GO:0009922 fatty acid elongase activity
906	GO:0102336 3-oxo-arachidoyl-CoA synthase activity
891	GO:0005432 calcium:sodium antiporter activity
874	GO:0015248 sterol transporter activity
829	GO:0140343 phosphatidylserine transfer activity
821	GO:0099580 monoatomic ion antiporter activity involved in regulation of postsynaptic membrane potential
810	GO:0005549 odorant binding
809	GO:0102338 3-oxo-lignoceronyl-CoA synthase activity
803	GO:0019841 retinol binding
801	GO:1905060 calcium:monoatomic cation antiporter activity involved in regulation of postsynaptic cytosolic calcium ion concentration
800	GO:0102337 3-oxo-cerotoyl-CoA synthase activity
778	GO:0052650 NADP-retinol dehydrogenase activity
774	GO:0102354 11-cis-retinol dehydrogenase activity
752	GO:0001618 virus receptor activity
681	GO:0005085 guanyl-nucleotide exchange factor activity

Table 3. Top 20 most frequent GO molecular functions in disease classes C1 and C2.

Transduction[42], *Olfactory transduction*[43], *and GPCR downstream signaling*[41]; all have literature evidence of cancer association.

Class C2-6 mainly consists of *Diseases of the eye and adnexa* (429/438) from a total of 527 eye diseases. We note that 18 eye diseases including cataract, ocular hypertension, and myopia, *Diseases of the eye and adnexa*, are in class C1. These 18 eye diseases have different disease drivers than the majority of eye diseases (class C2-6) and are closer to infectious diseases. Indeed, microbial infections can cause cataracts[44]. Class C2-6 has only 9 enriched pathways. Its topmost frequent pathway, the *Nuclear Receptor transcription pathway*, is associated with *glaucoma and retinoblastoma*[45]. The second ranked pathway is *Endocytosis*. Cells uptake exosomes by diverse forms of endocytosis, and exosomes play important roles in eye diseases[46]. The third and fifth ranked pathways are *Molecules associated with elastic fibres* and *Elastic fiber formation*. Elastic fibers are related to macular degeneration[47]. The 4th, *Integrin cell surface interactions,* is upregulated in *Retinitis Pigmentosa*[48].

Class C2-9-7 contains 265 diseases; 261 are *Diseases of the nervous system* including subtypes of *Alzheimer's disease* (AD) from a total 381 nervous system diseases. Note that a single GO function-GO:0003948:*N4-(beta-*

Frequency	Pathways
416 diseases of class C1	
416	REACT:R-HSA-198933 Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
416	REACT:R-HSA-168256 Immune System
416	REACT:R-HSA-1280218 Adaptive Immune System
415	REACT:R-HSA-449147 Signaling by Interleukins
415	REACT:R-HSA-168249 Innate Immune System
415	KEGG:hsa04640 Hematopoietic cell lineage
414	REACT:R-HSA-1280215 Cytokine Signaling in Immune system
412	KEGG:hsa05144 Malaria
412	KEGG:hsa04060 Cytokine-cytokine receptor interaction
410	REACT:R-HSA-1638074 Keratan sulfate/keratin metabolism
410	KEGG:hsa04630 Jak-STAT signaling pathway
410	KEGG:hsa04145 Phagosome
409	REACT:R-HSA-8854691 "Interleukin-19 20 22 24"
409	REACT:R-HSA-5602358 Diseases associated with the TLR signaling cascade
409	REACT:R-HSA-5260271 Diseases of Immune System
409	REACT:R-HSA-3560782 Diseases associated with glycosaminoglycan metabolism
409	REACT:R-HSA-2022854 Keratan sulfate biosynthesis
409	KEGG:hsa05152 Tuberculosis
409	KEGG:hsa04650 Natural killer cell mediated cytotoxicity
408	REACT:R-HSA-3656244 Defective B4GALT1 causes B4GALT1-CDG (CDG-2d)
3110 diseases of class C2	
1343	REACT:R-HSA-383280 Nuclear Receptor transcription pathway
1249	REACT:R-HSA-162582 Signal Transduction
1178	REACT:R-HSA-372790 Signaling by GPCR
1099	REACT:R-HSA-388396 GPCR downstream signaling
1068	KEGG:hsa04740 Olfactory transduction
1026	REACT:R-HSA-381753 Olfactory Signaling Pathway
900	KEGG:hsa_M00415 "Fatty acid biosynthesis elongation endoplasmic reticulum"
868	REACT:R-HSA-75876 Synthesis of very long-chain fatty acyl-CoAs
845	KEGG:hsa00062 Fatty acid elongation
796	REACT:R-HSA-216083 Integrin cell surface interactions
777	REACT:R-HSA-2453902 The canonical retinoid cycle in rods (twilight vision)
755	KEGG:hsa04144 Endocytosis
735	KEGG:hsa05412 Arrhythmogenic right ventricular cardiomyopathy (ARVC)
727	REACT:R-HSA-2129379 Molecules associated with elastic fibres
715	REACT:R-HSA-75105 Fatty Acyl-CoA Biosynthesis
705	KEGG:hsa05410 Hypertrophic cardiomyopathy (HCM)
701	KEGG:hsa05414 Dilated cardiomyopathy
660	REACT:R-HSA-1566948 Elastic fibre formation
658	REACT:R-HSA-2187338 Visual phototransduction
643	KEGG:hsa04512 ECM-receptor interaction

Table 4. Top 20 most frequent Reactome pathways in disease classes C1 and C2.

N-acetylglucosaminyl)-L-asparaginase activity involving AGA and ASRGL1 is present in 262 diseases. Protein L-asparaginase (ASRGL1) plays an important role in intercellular communication in neurodegenerative diseases[49]. The next most frequent pathway GO:0004364:*glutathione transferase activity* is only present in 20 diseases. Note that the topmost frequent pathway, *Glutathione conjugation,* is present only in 20 diseases. This is because for 241 diseases of this class, there is no significant (q-value < 0.05) pathway prediction. We thus relaxed the cutoff to q-value < 0.1 for this specific class only. The top pathway is still *Glutathione conjugation* with a frequency of 253 diseases (see Supplementary Table S8 (C2-9-7, q < 0.1). Increasing glutathione synthesis or glutathione conjugation activity can prevent neuronal cell loss[50]. The second to fourth most common pathways are *Regulation of PLK1 Activity at G2/M Transition*, *SCF-BTRC complex* and *Deactivation of the beta-catenin transactivating complex*, respectively, with>236≥ 236 appearances. PLK1 activity plays a critical role in neuronal autophagy[51] and is elevated in AD patients' brains[52]. The stem cell factor (SCF gene) can increase CXCR4 expression in therapies for Neurological Disorders[53], indicating that SCF is related to nervous system diseases. β-Catenin activity regulates dorsal–ventral pattern formation[54] and blood–brain barrier function in

Alzheimer's disease[55]. We note that these four pathways are predicted for all subtypes of AD (Alzheimer's disease 2,5,6,7,8,10,11,12,13,14,15) in our dataset.

Class C2-8, similar to C2-4, mainly contains *Neoplasms* (75/103) including 25 types of leukemia. The top 5 pathways are identical to those in class C2-4: *GPCR downstream signaling, Olfactory Signaling Pathway, Signaling by GPCR, Signal Transduction,* and *Olfactory transduction*; all have evidence of association with cancer. They differ after the top 5 pathways.

Class C2-9-12 consists of 124 diseases; 17 are *Neoplasms* including thyroid adenoma. 76 pathways are related to all 124 diseases. Here, *Signaling by the B Cell Receptor (BCR)* plays a role in Epstein–Barr virus infection causing malignancies[56]. *Interleukin receptor SHC signaling*, *Signaling by Insulin receptor* and the *Insulin receptor signaling cascade* are biomarkers for use of the anticancer drug Vinorelbine[57]. The *RET signaling* pathway is strongly correlated with increased risk of distant metastases[58]. The *PI3K/AKT Signaling* pathway plays a role in cancer onset and drug resistance[59]. The *RAF/MAP kinase cascade* is involved in TC21-mediated transformation of cell[60].

For the top 10 frequent pathways of the above 7 classes (C2-1, C2-2, C2-4, C2-6, C2-8, C2-9-7, C2-9-12), literature evidence indicates that their association with a given set of diseases is at least 77% correct. Apart from classes with more than 100 diseases that are dominated by a single clinical classification (>70%), there are also classes with a smaller numbers of diseases that are also dominated by a single clinical classification. Examples include class C2-9-1 with 6/6 of *Diseases of the circulatory system*, C2-9-2 with 5/7 for *Neoplasms*, C2-9-9 with 18/20 for *Congenital malformation, deformations and chromosomal abnormalitie* diseases, C2-9-13 with 3/4 for *Diseases of the circulatory system*, C2-9-15 and C2-9-16 with 8/11 and 13/16 for *Neoplasms*, respectively. *Neoplasms* dominate (>70%) 6 classes (C2-4:155/155, C2-8:75/103, C2-9-2:5/7, C2-9-12:117/124, C2-9-15:8/11, C2-9-16:13/16).

For the total 26 classes in Table S11, 17 have their top frequent pathways present in more than 90% of the diseases including classes found across clinical classifications. For example, class C2-3 has only 53/234 diseases of *Diseases of the musculoskeletal system and connective tissue* including *arthritis* and a few nervous diseases such as *epilepsy, migraine and multiple sclerosis*. The top 2 pathways, *Fatty acid biosynthesis elongation endoplasmic reticulum* and *Nuclear Receptor transcription pathway* are present in 229 and 227 diseases respectively. The fatty acid elongation pathway could be a potential target for treating *multiple sclerosis*[61]. The nuclear receptor activating transcription is expressed at elevated levels in inflamed joint tissues from patients with *arthritis*[62] and regulates seizure susceptibility in *epilepsy*[63].

Another example of a class with diverse clinical classifications is class C2-5 which has 43/124 diseases of *Neoplasms* including *brain cancer*, *bone marrow disease* and *stomach cancer*. The top 5 pathways, *Olfactory Signaling Pathway, Olfactory transduction, GPCR downstream signaling, Signal Transduction and Signaling by GPCR,* are all cancer related (see above C2-4, C2-8). Thus, the non-Neoplasm diseases, e.g. central nervous system lymphoma or aortic disease, in this class could have similar driving mechanisms as cancers.

A third example of diverse clinical classifications, particularly worthy of mention, is class C2-7 that includes 42/130 diseases of *Diseases of the genitourinary system* and 17/130 *metabolic diseases*. It contains 13 cancers including *cervical cancer*, *prostate cancer*, *liver cancer*, 9 nervous system diseases including *Alzheimer's disease*, *central nervous system disease*, 8 *Mental and behavioural disorders* including *eating disorder*, *nicotine dependence* and *opiate dependence*, 7 circulatory diseases including *atherosclerosis* and *coronary artery disease*, 6 digestive diseases including *Crohn's disease*, 5 respiratory diseases including *asthma* and *influenza*, and 5 eye diseases including *night blindness* and *central sleep apnea*. It is the closest class to C2-6 that is dominated by eye diseases (see Fig. 5). These apparently unrelated diseases are clustered together with their top 2 similar pathways, *the canonical retinoid cycle in rods (twilight vision) and Visual phototransduction,* having frequencies of 129 and 125, respectively. They both play a role in blindness[64,65]. Although their associations with other diseases mentioned above are not obvious, there are observed links between these diseases. *Asthma* has been reported to cause blindness[66]. *Crohn's disease* is comorbid with *asthma*[67], and *asthma* is a risk factor for *prostate cancer*[68]. Smokers with *asthma* have an increased odds of reporting high or very high propensity for *nicotine addiction*[69]. *Asthma* is also significantly associated with *coronary artery disease*[70].

What are the underlying mechanisms relating these diseases? Asthma is predicted to be related to *the canonical retinoid cycle in rods (twilight vision)* pathway through LRAT|RBP1|RBP3|RBP4|RDH11|RDH12|RDH8|RLBP1|RPE65. Prostate cancer is related through LRAT|RBP3|RDH11|RDH12|RDH8|RLBP1; Alzheimer's disease through LRAT|RBP3|RDH11|RDH12|RDH8|RLBP1|RPE65; and Crohn's disease, coronary artery disease and nicotine dependence through LRAT|RBP3|RBP4|RDH11|RDH12|RDH8|RLBP1|RPE65. All involve LRAT and a number of other proteins. There is literature evidence that the RBP4 gene in this pathway is involved in asthma and is a potential prognostic biomarker of non-allergic asthma caused by obesity in adolescents[71]. Its expression is decreased in the asthmatic lung compared to non-asthmatic controls[72]. Retinoid deficiency leads to airway hyperresponsiveness in a mice model[73]. Finally, LRAT is not expressed in prostate cancer cells[74]. It is even more significantly associated with Alzheimer's disease than APOE[75]. LRAT also impacts immune cells in Crohn's disease[76]. Coronary artery disease is affected by retinoic acid signaling involving LRAT[77]. RDH11 is associated with smoking behavior related to nicotine dependence[78]. Thus, LRAT, RBP4, RDH11 and their associated pathways are responsible for the connections between these diseases. Our predictions provide guidance for further investigations of the underlying mechanisms of these links. What is clear is that diseases of the eye are comorbid with many other diseases and examination of the eye might provide a noninvasive diagnostic approach for these diseases[79]. This study provides an explanation as to why this is the case.

Class C2-9-6 is the last example of a subclass with diverse clinical classifications. It involves 58 *Diseases of the respiratory system* including 9 subtypes of *pneumonia*, 11 *Diseases of the digestive system*, 10 *Diseases of the ear and mastoid process*, and 7 *Diseases of the genitourinary system*, and few other types of the 98 diseases. Despite

these diverse clinical classifications, its top pathway *Translocation of GLUT4 to the plasma membrane* is present in 92/98 diseases. Insulin stimulates the translocation of Glut4 to the plasma membrane[80], and increased insulin requirements are associated with pneumonia after severe injury[81]. This indicates that the *Translocation of GLUT4* pathway is related to pneumonia by insulin mediation. The second ranked pathway is *Tight junction* with a frequency of 90. It plays an important role in *Chlamydia pneumoniae* lung infection[82].

Literature evidence of clustered disease relationships

In addition to the above examples of literature evidence, to systematically show that the disease relationships within a cluster have supporting literature evidence and cannot be formed by random features that happened to be clustered, for each cluster we compiled a list of non-redundant disease pairs within a cluster and ranked them by their Euclidean distances (smallest ranked first). Non-redundant means diseases having identical predicted MOAs (e.g., disease subtypes) are considered to be one disease. A corresponding random pair list was constructed by replacing one disease of the pair with a disease randomly chosen from a different cluster. We then asked ChatGPT (https://platform.openai.com/) if the disease pairs are related based on literature evidence. For the 26 clusters and their top 10 ranked disease pairs, the average pairs of diseases returned as "yes" is 20%, whereas this average is only 4.2% for the corresponding randomized pairs. The enrichment factor is 4.8. Though 20% supporting literature evidence is not too high, it does not mean that the "novel" predictions are necessarily wrong. Rather, we plan on doing extensive text mining of clinical records to examine whether there is support for the unsubstantiated disease interrelationships. If we check the top 50, top 100 and top 500 instead of the top 10 ranked predictions, the average fraction of literature supported disease pairs are 15%, 13%, 11% respectively. In contrast, the average for randomized pairs are 5.0%, 5.3%, 5.4%, respectively, with corresponding enrichment factors of 3.0, 2.5, 2.0. Thus, the enrichment factors are correlated with the Euclidean distances of the disease pairs and the distance is biologically meaningful and cannot be attributed to random similar features (the randomized pairs have consistent ChatGPT evidence for 4–5% of the cases).

To further improve the robustness of this analysis, we leveraged an open-source literature review tool called Valsci. Valsci integrates SOTA large language models (in this case, gpt-4o) and uses chain-of-thought reasoning to determine the relevance, support, and contradicting evidence found in the Semantic Scholar[83] database of over 80 million academic papers and abstracts.

We ran approximately 1,700 disease pairs through Valsci. Half of the disease pairs were selected from the top 500 ranked pairs in clusters C1 and C2-1, and half were randomly paired diseases as discussed previously. Associations were tested in the form "[Condition X] and [Condition Y] share some overlapping pathogenic mechanisms that contribute to their respective disease processes." The tool output is a categorical score from Contradicted (0) to Highly Supported (5) for each claim. Associations lacking supporting or contradictory evidence receive a score of N/A (-1). For pairs within a disease cluster, 64 of 766 pairs have scores of 4 to 5, 702 pairs have score of − 1 to 3; for 928 randomized pairs, 40 pairs get scores of 4 to 5, 888 pairs get − 1 to 3. The calculated χ^2 statistic is 12 with a p-value of 0.0003. The rate of discovery for highly supported associations (score 4 to 5) improves from 4.3 to 8.4% for an enrichment factor of 1.9, roughly in line with the basic ChatGPT analysis.

Discussion

As indicated in Fig. 1 and its inset, the top 20 MOA proteins are involved in almost 70% of all diseases, with the top 100 MOA proteins involved on 80% of all diseases. Thus, there are remarkably few MOA proteins driving many diseases. This fact is responsible for the high extent of disease comorbidity; put another way, the disease network is remarkably dense. This has significant implications for the ability of selected drugs or immunotherapies to treat a broad spectrum of diseases.

As a consequence of the relatively few MOA proteins that are key disease drivers, we were able to classify 3526 diseases into 26 molecular classes after three levels of classification. 13 of 26 classes have a single clinical classification involving > 70% of the assigned diseases. At the top level, the majority (305/340) of infectious diseases are separated from the rest. Some clinically non-infectious diseases are also among this infection disease dominated class. These include common diseases such as constipation, colorectal cancer, large intestine cancer, muscular dystrophy, pneumonia, and bronchitis. We found that all top 5 frequent pathways of this class are immune system related (see Table 4). Not surprisingly, this characteristic implies that immune therapy/prevention could be a strategy for other diseases in this class. Non-infectious diseases are classified into 25 classes, 12 are dominated by a single clinical classification. But only one class, C2-4, is 100% pure Neoplasms. This paucity of MOA proteins likely is responsible for most observed disease comorbidities.

We then analyzed the GO molecular functions and Reactome pathways of each class and found that 17(21) of 26 classes have their top frequent pathways present in more than 90% (70%) of their member diseases including classes with mixed clinical classifications (see Supplementary Table S11). Consistent with the small number of MOA driver proteins, different clinical classifications share similar underlying molecular mechanisms. A literature search for evidence of predicted pathway associations shows a mean success rate of 77%. This allows us to find the molecular mechanism of diseases with unknown mechanisms with the help of known information from other members of the same class. One particularly interesting example is class C2-7 having diverse clinical classifications including common diseases such as *cervical cancer*, *prostate cancer*, *liver cancer*, *Alzheimer's disease*, *central nervous system disease*, *eating disorder*, *nicotine dependence*, *opiate dependence*, *atherosclerosis*, *coronary artery disease*, *Crohn's disease*, *asthma*, and *night blindness*. We found that these apparently unrelated diseases share the same pathway: *the canonical retinoid cycle in rods (twilight vision)*. This pathway is obviously related to eye diseases. Indeed, this class is one of the closest classes to class C2-6 dominated by eye diseases (Fig. 5). We found the underlying proteins linking these diseases are LRAT, RDH11, RBP4 and possibly others. As indicated above, the eye could be a powerful tool to help diagnose serious comorbid diseases. The shared

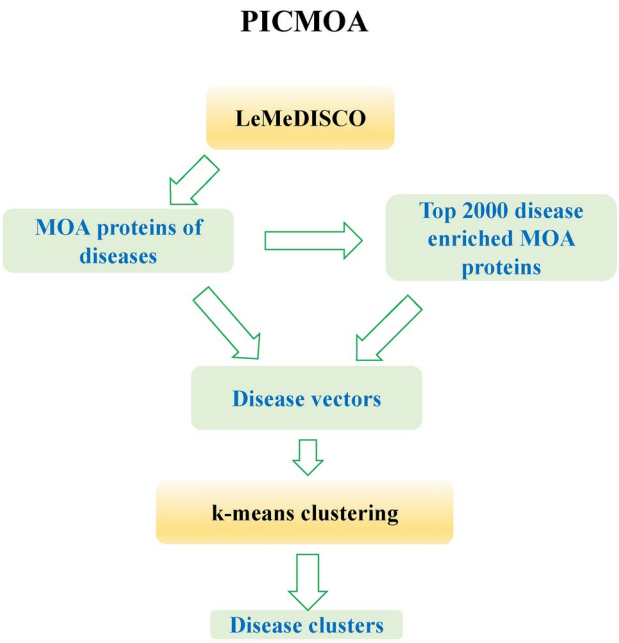

Fig. 6. Schematic diagram of PICMOA method.

pathways reflecting common MOAs of diseases in the same class can facilitate drug discovery. By targeting these shared pathways, drugs might be developed for a class of diseases or repositioned from one member of the class to other members. Future studies will employ these disease networks to predict "broad spectrum" repurposed drugs that might treat an entire set of related diseases.

PICMOA, which classifies complex diseases with many MOA proteins, can be extended to classify rare Mendelian diseases that are usually caused by mutations in a single gene or a few genes. In order to do the classification for rare diseases similar to PICMOA, one needs to have a set of genes that cover the majority of rare diseases. Of the 2000 basis genes employed in PICMOA, only 508 (p-value = 3.88×10^{-11}) genes overlap with the 6095 unique genes in the OMIM database[15]. Thus, to classify both complex and rare diseases, the number of genes will need to be extended to include the 6095 unique OMIM genes. This issue will also be explored in our future work.

Despite the successful prediction of pathways of each class and in previous work, comorbidities[2], there are limitations of the current work. The MOA proteins associated with each disease are derived from known and predicted efficacious drugs for an average patient, and thus the approach is not yet ready for personalized medicine/treatment. To tackle this issue, mapping an individual patient's genetic information (i.e., exome single nucleotide variations-SNP) to disease classes is needed; in that regard, work is underway to develop a more personalized approach to precision medicine and diagnostics.

Materials and methods

A schematic diagram of PICMOA is shown in Fig. 6. We detail each of the steps below.

Disease enriched mode of action proteins

For each pair of the 3526 diseases in our library, in LeMeDISCO[2] we used MEDICASCY[22] to derive the associated MOA proteins based on their p-value (which is then converted to a q-value, with a q-value cutoff of 0.05)[83]. To further prioritize them according to their likelihood of contributing to efficacy, the p-value weighted frequency of shared MOA proteins across the top 100 predicted comorbidities is calculated. We define a p-value weighted frequency of an input MOA protein T as follows: If protein T is shared by a comorbid disease D and the p-value of T associated with D is P, then its comorbid disease weight defined by the min (1.0,− αlogP) is counted as T's frequency. In practice, we used data from 10 cancer cell lines[84] to determine the coefficient α to 0.025 for anticancer drug discovery. The idea is to use this information to predict which are the driver proteins for a given cancer and then use this information to predict the most likely effective drugs. By varying the parameter α from 0.005 to 0.25 on these ten cell lines, we set α = 0.025 which gives the best CoVLS drug screening[85] mean success rate of 86%. Further details are found in Supplementary Materials.

Here, we extend the above procedure for prioritizing individual disease MOA proteins to prioritize all human proteins in all 3526 diseases to predict their sets of comorbid diseases. For each human protein, we calculate the p-value weighted frequency among the 3526 diseases. This results in a ranked list of all human proteins (a total 14,683 unique genes are ranked with non-zero frequency) according to their likelihood of being efficacious MOA proteins for a set of diseases, rather than being only comorbid to a specific disease. The frequencies are also normalized by the number of diseases considered (here 3526). The top 2000 proteins comprise the MOA protein vector of a given disease.

Classification of diseases using disease enriched MOA proteins

To classify diseases, we vectorize them by projecting the predicted associated MOA proteins of individual disease to the top 2000 genes with the ranked normalized frequencies as the vector components. We then cluster the diseases by k-means clustering using the "k-means" R-package function with 1000 maximum iterations and 25 random sets similar to what was done in Ref.[7] for cancer patients. Here, the optimal number of clusters is not fixed (as opposed to Ref.[7] that sets it to 10), but is determined by the Silhouette Coefficient[23] and is obtained by scanning a range of cluster numbers from 2 to 50. The optimal number of clusters is the one having the peak Silhouette Coefficient[23].

Function and pathway annotation of molecular disease classes

The MOA proteins represented by the vector of a given disease are used to search for the molecular functional enrichment of the given disease using the GO data set[29,30] downloaded from http://geneontology.org/docs/download-ontology/. Pathways are computed using the Reactome data set[31]. For a given disease and its list of MOA proteins from its vector (i.e. only a subset of its full list are considered, MOA proteins outside of the top 2000 base proteins are ignored), we calculated the p-value of its MOA proteins that overlap with each GO function or Reactome pathway using Fisher's exact test[86]. Then, the p-values were converted to q-values. Here, the q-value is calculated by an empirical equation following the Benjamini–Hochberg procedure[87].

$$q-value = p-value \times \frac{total\ number\ of\ pathways\ or\ GO\ functions\ tested}{rank\ of\ the\ p-value} \quad (1)$$

Ranking using the p-value criterion involves sorted p-values, with smallest ranked first.

An enriched GO function or Reactome pathway is defined when the q-value of the association is < 0.05[83].

Data availability

All data presented in this work are freely available at https://sites.gatech.edu/cssb/PICMOA/. And data and scripts for reproducing the results can be found at https://github.com/hzhou3ga/picmoa.

Received: 3 June 2024; Accepted: 6 March 2025
Published online: 20 March 2025

References

1. Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms–disease network. *Nat. Commun.* **5**, 4212 (2014).
2. Courtney, A., Zhou, H., Ilkowski, B., Forness, J. & Skolnick, J. LeMeDISCO is a computational method for large-scale prediction & molecular interpretation of disease comorbidity. *Commun. Biol.* **5**, 870 (2022).
3. Ko, Y., Cho, M., Lee, J.-S. & Kim, J. Identification of disease comorbidity through hidden molecular mechanisms. *Sci. Rep.* **6**, 39433 (2016).
4. Anonymous. https://www.medicalbillingandcoding.org/icd-10-cm/.
5. Perou, C. M. et al. Molecular portraits of human breast tumours. *Nature* **406**, 747–752. https://doi.org/10.1038/35021093 (2000).
6. Chang, K. et al. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* **45**, 1113–1120. https://doi.org/10.1038/ng.2764 (2013).
7. Chen, F. et al. Pan-cancer molecular classes transcending tumor lineage across 32 cancer types, multiple data platforms, and over 10,000 cases. *Clin. Cancer Res.* **24**, 2182 (2018).
8. Zhang, Y. et al. A pan-cancer proteogenomic atlas of PI3K/AKT/mTOR pathway alterations. *Cancer Cell* **31**, 820–832. https://doi.org/10.1016/j.ccell.2017.04.013 (2017).
9. Oti, M. & Brunner, H. The modular nature of genetic diseases. *Clin. Genet.* **71**, 1–11. https://doi.org/10.1111/j.1399-0004.2006.00708.x (2007).
10. Schriml, L. M. et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962. https://doi.org/10.1093/nar/gky1032 (2019).
11. Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci.* **99**, 6567–6572. https://doi.org/10.1073/pnas.082099299 (2002).
12. Reinius, L. E. et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PLoS One* **7**, e41361. https://doi.org/10.1371/journal.pone.0041361 (2012).
13. Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat. Rev. Genet.* **16**, 172–183. https://doi.org/10.1038/nrg3871 (2015).
14. Oketch, D. J. A., Giulietti, M. & Piva, F. Copy number variations in pancreatic cancer: from biological significance to clinical utility. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms25010391 (2023).
15. *Online Mendelian Inheritance in Man, OMIM®*, https://omim.org/.
16. Melamed, R. D., Emmett, K. J., Madubata, C., Rzhetsky, A. & Rabadan, R. Genetic similarity between cancers and comorbid Mendelian diseases identifies candidate driver genes. *Nat. Commun.* **6**, 7033. https://doi.org/10.1038/ncomms8033 (2015).
17. World Health. *O* (World Health Organization, 2004).
18. The UniProt, C. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489. https://doi.org/10.1093/nar/gkaa1100 (2021).
19. Klug, A. The discovery of zinc fingers and their applications in gene regulation and genome manipulation. *Annu. Rev. Biochem.* **79**, 213–231. https://doi.org/10.1146/annurev-biochem-010909-095056 (2010).
20. Stelzer, G. et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. *Curr. Protoc. Bioinform.* **54**, 13031–313033. https://doi.org/10.1002/cpbi.5 (2016).
21. Sayers, E. W. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20-d26. https://doi.org/10.1093/nar/gkab1112 (2022).
22. Zhou, H. et al. MEDICASCY: A machine learning approach for predicting small-molecule drug side effects, indications, efficacy, and modes of action. *Mol. Pharm.* **17**, 1558–1574. https://doi.org/10.1021/acs.molpharmaceut.9b01248 (2020).
23. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7 (1987).
24. Thorndike, R. L. Who Belongs in the Family?. *Psychometrika* **18**, 267–276. https://doi.org/10.1007/BF02289263 (1953).

25. Zhao, Y. & Yu, Y. B. Intestinal microbiota and chronic constipation. *Springerplus* **5**, 1130. https://doi.org/10.1186/s40064-016-2821-1 (2016).
26. Panwalker, A. P. Unusual infections associated with colorectal cancer. *Rev. Infect. Dis.* **10**, 347–364. https://doi.org/10.1093/clinids/10.2.347 (1988).
27. Naveed, Z. et al. Association of COVID-19 infection with incident diabetes. *JAMA Netw. Open* **6**, e238866–e238866. https://doi.org/10.1001/jamanetworkopen.2023.8866 (2023).
28. Tamura, K., Stecher, G. & Kumar, S. MEGA11: Molecular evolutionary genetics analysis version 11. *Mol. Biol. Evol.* **38**, 3022–3027. https://doi.org/10.1093/molbev/msab120 (2021).
29. Ashburner, M. et al. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29. https://doi.org/10.1038/75556 (2000).
30. The Gene Ontology, C. et al. The gene ontology knowledgebase in 2023. *Genetics* **224**, 031. https://doi.org/10.1093/genetics/iyad031 (2023).
31. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498-d503. https://doi.org/10.1093/nar/gkz1031 (2020).
32. Tu, X., Huang, H., Xu, S., Li, C. & Luo, S. Single-cell transcriptomics reveals immune infiltrate in sepsis. *Front. Pharmacol.* **14**, 1133145. https://doi.org/10.3389/fphar.2023.1133145 (2023).
33. Ruck, C., Reikie, B. A., Marchant, A., Kollmann, T. R. & Kakkar, F. Linking susceptibility to infectious diseases to immune system abnormalities among HIV-exposed uninfected infants. *Front. Immunol.* https://doi.org/10.3389/fimmu.2016.00310 (2016).
34. Janeway, C. A. How the immune system works to protect the host from infection: A personal view. *Proc. Natl. Acad. Sci.* **98**, 7461–7468. https://doi.org/10.1073/pnas.131202998 (2001).
35. Mujalli, A. et al. Bioinformatics insights into the genes and pathways on severe COVID-19 pathology in patients with comorbidities. *Front. Physiol.* **13**, 1045469. https://doi.org/10.3389/fphys.2022.1045469 (2022).
36. Dagenais, A., Villalba-Guerrero, C. & Olivier, M. Trained immunity: A "new" weapon in the fight against infectious diseases. *Front. Immunol.* https://doi.org/10.3389/fimmu.2023.1147476 (2023).
37. Fatemifar, G. et al. Genome-wide association study of primary tooth eruption identifies pleiotropic loci associated with height and craniofacial distances. *Hum. Mol. Genet.* **22**, 3807–3817. https://doi.org/10.1093/hmg/ddt231 (2013).
38. Gao, X. et al. Chronic stress promotes colitis by disturbing the gut microbiota and triggering immune system response. *Proc. Natl. Acad. Sci.* **115**, E2960–E2969. https://doi.org/10.1073/pnas.1720696115 (2018).
39. Zhang, J. et al. Lipid metabolism in type 1 diabetes mellitus: Pathogenetic and therapeutic implications. *Front. Immunol.* **13**, 999108. https://doi.org/10.3389/fimmu.2022.999108 (2022).
40. Weber, L. et al. Characterization of the olfactory receptor OR10H1 in human urinary bladder cancer. *Front. Physiol.* **9**, 456 (2018).
41. Chaudhary, P. K. & Kim, S. An insight into GPCR and G-proteins as cancer drivers. *Cells* https://doi.org/10.3390/cells10123288 (2021).
42. Sever, R. & Brugge, J. S. Signal transduction in cancer. *Cold Spring Harb. Perspect. Med.* https://doi.org/10.1101/cshperspect.a006098 (2015).
43. Sanz, G. L. I. et al. Promotion of cancer cell invasiveness and metastasis emergence caused by olfactory receptor stimulation. *PLoS One* **9**, e85110 (2014).
44. Lotti, R. & Dart, J. K. Cataract as a complication of severe microbial keratitis. *Eye (Lond)* **6**(Pt 4), 400–403. https://doi.org/10.1038/eye.1992.82 (1992).
45. Ahmad, B., Leila, S. & Taleahmad, S. Investigation of key signaling pathways associating miR-204 and common retinopathies. *BioMed Res. Int.* **2021**, 5568113 (2021).
46. Liu, J. et al. Roles of exosomes in ocular diseases. *Int. J. Nanomed.* **15**, 10519–10538. https://doi.org/10.2147/ijn.S277190 (2020).
47. Nita, M., Strzałka-Mrozik, B., Grzybowski, A., Mazurek, U. & Romaniuk, W. Age-related macular degeneration and changes in the extracellular matrix. *Med. Sci. Monit.* **20**, 1003–1016. https://doi.org/10.12659/msm.889887 (2014).
48. Bielmeier, C. B. et al. Transcriptional profiling identifies upregulation of neuroprotective pathways in retinitis pigmentosa. *Int. J. Mol. Sci.* https://doi.org/10.3390/ijms22126307 (2021).
49. Huo, L., Du, X., Li, X., Liu, S. & Xu, Y. The emerging role of neural cell-derived exosomes in intercellular communication in health and neurodegenerative diseases. *Front. Neurosci.* **15**, 738442 (2021).
50. Hirth, F. Drosophila melanogaster in the study of human neurodegeneration. *CNS Neurol. Disord. Drug Targets.* **9**, 504–523. https://doi.org/10.2174/187152710791556104 (2010).
51. Chen, L.-L. et al. Phosphoproteome-based kinase activity profiling reveals the critical role of MAP2K2 and PLK1 in neuronal autophagy. *Autophagy* **13**, 1969–1980. https://doi.org/10.1080/15548627.2017.1371393 (2017).
52. Song, B. et al. Inhibition of Polo-like kinase 1 reduces beta-amyloid-induced neuronal cell death in Alzheimer's disease. *Aging (Albany NY).* **3**, 846–851. https://doi.org/10.18632/aging.100382 (2011).
53. Andrzejewska, A., Dabrowska, S., Lukomska, B. & Janowski, M. Mesenchymal stem cells for neurological disorders. *Adv. Sci. (Weinh)* **8**, 2002944. https://doi.org/10.1002/advs.202002944 (2021).
54. Liu, J.-X. et al. Eaf1 and Eaf2 mediate zebrafish dorsal-ventral axis patterning via suppressing Wnt/β-Catenin activity. *Int. J. Biol. Sci.* **14**, 705–716. https://doi.org/10.7150/ijbs.18997 (2018).
55. Wang, Q. et al. Activation of Wnt/β-catenin pathway mitigates blood-brain barrier dysfunction in Alzheimer's disease. *Brain* **145**, 4474–4488. https://doi.org/10.1093/brain/awac236 (2022).
56. Fish, K. et al. Rewiring of B cell receptor signaling by Epstein-Barr virus LMP2A. *Proc. Natl. Acad. Sci.* **117**, 26318–26327. https://doi.org/10.1073/pnas.2007946117 (2020).
57. Sachdev, P., Ronen, R., Dutkowski, J. & Littlefield, B. A. Systematic analysis of genetic and pathway determinants of eribulin sensitivity across 100 human cancer cell lines from the cancer cell line encyclopedia (CCLE). *Cancers (Basel)* https://doi.org/10.3390/cancers14184532 (2022).
58. Regua, A. T., Najjar, M. & Lo, H.-W. RET signaling pathway and RET inhibitors in human cancer. *Front. Oncol.* https://doi.org/10.3389/fonc.2022.932353 (2022).
59. Rascio, F. et al. The pathogenic role of PI3K/AKT pathway in cancer onset and drug resistance: an updated review. *Cancers (Basel).* https://doi.org/10.3390/cancers13163949 (2021).
60. Rosário, M., Paterson, H. F. & Marshall, C. J. Activation of the Raf/MAP kinase cascade by the Ras-related protein TC21 is required for the TC21-mediated transformation of NIH 3T3 cells. *Embo J.* **18**, 1270–1279. https://doi.org/10.1093/emboj/18.5.1270 (1999).
61. Garcia Corrales, A. V. et al. Fatty acid elongation by ELOVL6 hampers remyelination by promoting inflammatory foam cell formation during demyelination. *Proc. Natl. Acad. Sci.* **120**, e2301030120. https://doi.org/10.1073/pnas.2301030120 (2023).
62. McCoy, J. M. et al. Orphan nuclear receptor NR4A2 induces transcription of the immunomodulatory peptide hormone prolactin. *J. Inflamm. (Lond).* **12**, 13. https://doi.org/10.1186/s12950-015-0059-2 (2015).
63. Pönniö, T. & Conneely, O. M. nor-1 regulates hippocampal axon guidance, pyramidal cell survival, and seizure susceptibility. *Mol. Cell Biol.* **24**, 9070–9078. https://doi.org/10.1128/mcb.24.20.9070-9078.2004 (2004).
64. Ali, M. U., Rahman, M. S. U., Cao, J. & Yuan, P. X. Genetic characterization and disease mechanism of retinitis pigmentosa; current scenario. *3 Biotech* **7**, 251. https://doi.org/10.1007/s13205-017-0878-3 (2017).
65. Mustafi, D., Arbabi, A., Ameri, H. & Palczewski, K. Retinal gene distribution and functionality implicated in inherited retinal degenerations can reveal disease-relevant pathways for pharmacologic intervention. *Pharmaceuticals (Basel).* https://doi.org/10.3390/ph12020074 (2019).

66. Kunishige, T., Omori, A., Tateno, A., Yahata, N. & Hori, J. Cortical blindness caused by hypoxemia following an asthma attack. *Jpn. J. Ophthalmol.* **55**, 588–590. https://doi.org/10.1007/s10384-011-0058-7 (2011).
67. Kuenzig, M. E., Bishay, K., Leigh, R., Kaplan, G. G. & Benchimol, E. I. Co-occurrence of asthma and the inflammatory bowel diseases: a systematic review and meta-analysis. *Clin. Transl. Gastroenterol.* **9**, 188. https://doi.org/10.1038/s41424-018-0054-z (2018).
68. Su, Y. L., Chou, C. L., Rau, K. M. & Lee, C. T. Asthma and risk of prostate cancer: a population-based case-cohort study in Taiwan. *Medicine (Baltimore)* **94**, e1371. https://doi.org/10.1097/md.0000000000001371 (2015).
69. Perret, J. L., Bonevski, B., McDonald, C. F. & Abramson, M. J. Smoking cessation strategies for patients with asthma: improving patient outcomes. *J. Asthma Allergy* **9**, 117–128. https://doi.org/10.2147/jaa.S85615 (2016).
70. Wang, L., Gao, S., Yu, M., Sheng, Z. & Tan, W. Association of asthma with coronary heart disease: A meta analysis of 11 trials. *PLoS One* **12**, e0179335. https://doi.org/10.1371/journal.pone.0179335 (2017).
71. Leija-Martínez, J. J. *et al.* Retinol-binding protein 4 and plasminogen activator inhibitor-1 as potential prognostic biomarkers of non-allergic asthma caused by obesity in adolescents *Allergologia et Immunopathologia* **49**, 21–29 (2021).
72. Defnet, A. E. et al. Dysregulated retinoic acid signaling in airway smooth muscle cells in asthma. *Faseb J* **35**, e22016. https://doi.org/10.1096/fj.202100835R (2021).
73. Chen, F. et al. Prenatal retinoid deficiency leads to airway hyperresponsiveness in adult mice. *J. Clin. Investig.* **124**, 801–811. https://doi.org/10.1172/jci70291 (2014).
74. Guo, X. et al. Retinol metabolism and lecithin:retinol acyltransferase levels are reduced in cultured human prostate cancer cells and tissue specimens. *Cancer Res.* **62**, 1654–1661 (2002).
75. Ertekin-Taner, N. Genetics of Alzheimer disease in the pre- and post-GWAS era. *Alzheimer's Res. Ther.* **2**, 3. https://doi.org/10.1186/alzrt26 (2010).
76. Oliveira, L. D. M., Teixeira, F. M. E. & Sato, M. N. Impact of retinoic acid on immune cells and inflammatory diseases. *Mediat. Inflamm.* **2018**, 3067126. https://doi.org/10.1155/2018/3067126 (2018).
77. Wang, S. & Moise, A. R. Recent insights on the role and regulation of retinoic acid signaling during epicardial development. *Genesis* **57**, e23303. https://doi.org/10.1002/dvg.23303 (2019).
78. Han, S., Gelernter, J., Luo, X. & Yang, B. Z. Meta-analysis of 15 genome-wide linkage scans of smoking behavior. *Biol. Psychiatry* **67**, 12–19. https://doi.org/10.1016/j.biopsych.2009.08.028 (2010).
79. Mukamal, R. *20 Surprising Health Problems an Eye Exam Can Catch*. https://www.aao.org/eye-health/tips-prevention/surprising-health-conditions-eye-exam-detects (2023).
80. Huang, P., Altshuller, Y. M., Hou, J. C., Pessin, J. E. & Frohman, M. A. Insulin-stimulated plasma membrane fusion of Glut4 glucose transporter-containing vesicles is regulated by phospholipase D1. *Mol. Biol. Cell* **16**, 2614–2623. https://doi.org/10.1091/mbc.e04-12-1124 (2005).
81. Martin, R. S. et al. Increased insulin requirements are associated with pneumonia after severe injury. *J. Trauma* **63**, 358–364. https://doi.org/10.1097/TA.0b013e31809ed905 (2007).
82. Chiba, N. et al. Mast cells play an important role in chlamydia pneumoniae lung infection by facilitating immune cell recruitment into the airway. *J. Immunol.* **194**, 3840–3851. https://doi.org/10.4049/jimmunol.1402685 (2015).
83. John, D. S. The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035. https://doi.org/10.1214/aos/1074290335 (2003).
84. NCI-60 Human Tumor Cell Lines Screen: https://dtp.cancer.gov/discovery_development/nci-60/.
85. Astore, C., Zhou, H., Jacob, J. & Skolnick, J. Prediction of severe adverse events, modes of action and drug treatments for COVID-19's complications. *Sci. Rep.* **11**, 20864. https://doi.org/10.1038/s41598-021-00368-6 (2021).
86. Fisher, R. A. On the interpretation of χ 2 from contingency tables, and the calculation of P. *J. R. Stat. Soc.* **85**, 87–94 (1922).
87. Agresti, A. *Categorical Data Analysis* (Wiley, 1990).

Acknowledgements

We thank Bartosz Ilkowski for computing support and Jessica Forness for proof-reading the manuscript.

Author contributions

Conceptualization: J.S., H.Z. Methodology: H.Z., J.S. Investigation: H.Z., B.E., J.S. Funding acquisition: J.S. Project administration: J.S. Supervision: J.S. Writing—original draft: H.Z. Writing—review and editing: J.S., H.Z., B.E.

Funding

This project was funded by R35GM-118039 of the Division of General Medical Sciences of the NIH.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version that contains supplementary material is available at https://doi.org/10.1038/s41598-025-93377-8.

Correspondence and requests for materials should be addressed to J.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.